Variable Span Filters for Speech Enhancement

Jensen, Jesper Rindom; Benesty, Jacob; Christensen, Mads Græsbøll

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2016.7472930

Publication date:
2016

Document Version
Accepted author manuscript, peer reviewed version

VARIABLE SPAN FILTERS FOR SPEECH ENHANCEMENT

Jesper Rindom Jensen[1]*, Jacob Benesty*[2]*, and Mads Græsbøll Christensen*[1]

[1]Audio Analysis Lab, AD:MT, Aalborg University, Aalborg, Denmark, {jrj,mgc}@create.aau.dk
[2]INRS-EMT, University of Quebec, Montreal, Canada, benesty@emt.inrs.ca

ABSTRACT

In this work, we consider enhancement of multichannel speech recordings. Linear filtering and subspace approaches have been considered previously for solving the problem. The current linear filtering methods, although many variants exist, have limited control of noise reduction and speech distortion. Subspace approaches, on the other hand, can potentially yield better control by filtering in the eigen-domain, but traditionally these approaches have not been optimized explicitly for traditional noise reduction and signal distortion measures. Herein, we combine these approaches by deriving optimal filters using a joint diagonalization as a basis. This gives excellent control over the performance, as we can optimize for noise reduction or signal distortion performance. Results from real data experiments show that the proposed variable span filters can achieve better performance than existing filters. In terms of output SNR, the gain was more than 8 dB, and more than 0.1 in mean opinion score in the conducted experiments.

Index Terms— Speech enhancement, joint diagonalization, optimal filtering, multichannel enhancement, tradeoff filter

1. INTRODUCTION

Noise reduction, or speech enhancement, is an essential tool in many important applications, including hearing aids, teleconferencing, and mobile telephony. The topic has, therefore, attracted a lot of attention, resulting in many different speech enhancements methods. Roughly, these can be categorized into linear filtering methods [1], spectral subtractive methods [2], statistical methods [3–5], and subspace methods [6, 7]. These works are all on single-channel speech enhancement, while multichannel enhancement, which is the topic of this paper, has also been attracting considerable attention (see, e.g., [8–11]). We refer the interested reader to [1, 12, 13] for overviews of recent advances in noise reduction.

While many different approaches to speech enhancement have been considered, we here consider the linear filtering and subspace approaches. In the methods based on linear filtering, noise reduction is obtained by convolution of the observed signal, which comprises both the signal of interest and the additive noise, with the impulse response of a filter. The noise reduction problem then amounts to designing this filter so that it meets some requirements, in terms of, for example, noise reduction and speech distortion. For example, when the mean-square error (MSE) is used as a performance measure and the filter is optimized so as to minimize the MSE, the classical Wiener filter is obtained. In subspace methods [14, 15], a diagonalization of the involved correlation matrices is obtained by means of, for example, the Karhunen-Loève transform, the eigenvalue decomposition, or the singular value decomposition, and this is then used for noise reduction by identifying bases for the speech-plus-noise subspace (also sometimes simply called the signal subspace) and the noise subspace, respectively. Of particular relevance to the present work, is the prior use of joint diagonalization for noise reduction, something that has previously been done in [7] and later in [8, 16] to account for colored noise.

In this paper, we express the multichannel noise reduction problem as a linear filtering problem using joint diagonalization of the correlation matrix of the signal of interest and the noise. More specifically, we consider filter designs, wherein the filter coefficients are formed as linear combinations of a desired number of eigenvectors. This way, speech distortion can be traded for more noise reduction in a simple way by changing the number of eigenvectors. We also proposed enhancement filters based on the joint diagonalization in [17, 18], but these were only derived for the single-channel case. Moreover, these enhancement methods used an indirect approach where the noise is estimated first and subtracted from the observation to obtain the enhanced signal, whereas the filters proposed herein estimates the desired signal directly. In the proposed framework, a number of noise reduction filters, which are referred to as variable span filters, are derived. These include maximum SNR, minimum distortion, Wiener, and tradeoff filters. Compared to previous subspace methods, this enable us to quantify, and optimize for, the noise reduction and speech distortion performances.

The remainder of the paper is organized as follows: in Section 2, the signal model and problem formulation are presented. Then, in Section 3, the variable span filters are proposed. Finally, we present some experimental results in Section 4 and conclude on the work in Section 5.

2. NOISE REDUCTION PROBLEM

We consider the scenario where we have an array of microphones, consisting of M sensors, that captures a sound field containing a desired speech source as well as noise. This usual, multichannel signal model [19, 20] can also be written as

$$y_m(t) = g_m(t) * s(t) + v_m(t) = x_m(t) + v_m(t), \qquad (1)$$

for $m = 1, 2, \ldots, M$, where $(\cdot)_m$ denotes a variable associated with sensor m, $y_m(t)$ is the observed signal, $g_m(t)$ is the room impulse response from the source to sensor m, and $v_m(t)$ is the undesired noise. Furthermore, we introduced an additional variable, $x_m(t) = g_m(t) * s(t)$, which will be treated as the desired speech signal, since we do not consider the dereverberation problem herein.

If we then apply the short-time Fourier transform (STFT) on the observed signal, we instead get a time-frequency domain model:

$$Y_m(k,n) = X_m(k,n) + V_m(k,n),\ m = 1, 2, \ldots, M, \qquad (2)$$

with $Y_m(k,n)$, $X_m(k,n)$, and $V_m(k,n)$ being the STFTs of $y_m(t)$, $x_m(t)$, and $v_m(t)$, respectively at frequency bin $k \in \{0, 1, \ldots, K-$

This work was supported by the Danish Council for Independent Research, grant ID: DFF 1337-00084, and the Villum foundation.

$1\}$ and time frame n. That is, each of these variables are zero-mean and complex. To facilitate the derivation of noise reduction methods, we introduce a more convenient vector model as

$$\mathbf{y}(k,n) = [Y_1(k,n) \ \cdots \ Y_M(k,n)]^T = \mathbf{x}(k,n) + \mathbf{v}(k,n),$$

where we have defined $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ similarly to $\mathbf{y}(k,n)$. If we assume that the desired speech, $X_m(k,n)$, and the noise, $V_m(k,n)$, are uncorrelated, we can write the correlation matrix of $\mathbf{y}(k,n)$ as

$$\mathbf{\Phi}_{\mathbf{y}}(k,n) = E\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right] = \mathbf{\Phi}_{\mathbf{x}}(k,n) + \mathbf{\Phi}_{\mathbf{v}}(k,n),$$

with $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ being the correlation matrices of $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$, respectively.

A general assumption in many multichannel, speech enhancement methods operating in the STFT domain is that $X_m(k,n) = G_m(k)S(k,n)$, for $m = 1,2,\ldots,M$, where $G_m(k)$ and $S(k,n)$ are the STFTs of $g_m(t)$ and $s(t)$, respectively. Clearly, the rank of $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ is equal to 1 when this assumption holds. The assumption, however, is only valid when the analysis window of the STFT is infinitely long, but this is obviously never the case in practice [21]. As a consequence, the rank of $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ will instead be a positive integer between 1 and M. Studying noise reduction algorithms while taking this fact into account is, therefore, of great interest and is one of the contributions of this paper.

Since short time windows are often preferred in the computation of the STFT, there will inevitably be some correlation between consecutive time frames. We take this interframe correlation into account in the filter designs in Section 3. To achieve this, we consider N consecutive frames, and rewrite the observations as

$$\begin{aligned}\underline{\mathbf{y}}(k,n) &= \left[\mathbf{y}^T(k,n)\ \mathbf{y}^T(k,n-1)\ \cdots\ \mathbf{y}^T(k,n-N+1)\right]^T \\ &= \underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n),\end{aligned} \tag{3}$$

where $\underline{\mathbf{x}}(k,n)$ and $\underline{\mathbf{v}}(k,n)$ are defined similarly to $\mathbf{y}(k,n)$. The correlation matrix of the stacked observations, $\underline{\mathbf{y}}(k,n)$, is then

$$\mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n) = E\left[\underline{\mathbf{y}}(k,n)\underline{\mathbf{y}}^H(k,n)\right] = \mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n) + \mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n),$$

with $\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)$ and $\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$ being the correlation matrices of $\underline{\mathbf{x}}(k,n)$ and $\underline{\mathbf{v}}(k,n)$, respectively. Moreover, the rank of these are assumed to be equal to $P < MN$ and MN.

If we chose sensor 1 as our reference sensor, the multichannel, speech enhancement problem in the STFT domain is then to recover $X_1(k,n)$ from the observations $\underline{\mathbf{y}}(k,n)$ as well as possible. That is, we should have large degree of noise reduction and only little distortion of the desired signal.

To approach the speech enhancement problem, we first consider a joint diagonalization of the signal and noise correlation matrices [22]:

$$\mathbf{B}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\mathbf{B}(k,n) = \mathbf{\Lambda}(k,n), \tag{4}$$

$$\mathbf{B}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)\mathbf{B}(k,n) = \mathbf{I}_{MN}, \tag{5}$$

with $\mathbf{B}(k,n)$ being a full-rank square matrix (of size $MN \times MN$), $\mathbf{\Lambda}(k,n)$ a diagonal matrix whose main elements are real and non-negative, and $\mathbf{I}_{MN}$ the $MN \times MN$ identity matrix. Moreover, the matrices $\mathbf{\Lambda}(k,n)$ and $\mathbf{B}(k,n)$ are the eigenvalue and -vector matrices, respectively, of $\mathbf{\Phi}_{\underline{\mathbf{v}}}^{-1}(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)$, i.e.,

$$\mathbf{\Phi}_{\underline{\mathbf{v}}}^{-1}(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\mathbf{B}(k,n) = \mathbf{B}(k,n)\mathbf{\Lambda}(k,n). \tag{6}$$

The rank of the matrix $\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)$ is assumed to be equal to P, so the eigenvalues of $\mathbf{\Phi}_{\underline{\mathbf{v}}}^{-1}(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)$ can be ordered as $\lambda_1(k,n) \geq \lambda_2(k,n) \geq \cdots \geq \lambda_P(k,n) > \lambda_{P+1}(k,n) = \cdots = \lambda_{MN}(k,n) = 0$. That is, the first P and last $MN-P$ eigenvalues of the matrix product $\mathbf{\Phi}_{\underline{\mathbf{v}}}^{-1}(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)$ are positive and exactly zero, respectively. The vectors $\underline{\mathbf{b}}_1(k,n), \underline{\mathbf{b}}_2(k,n), \ldots, \underline{\mathbf{b}}_{MN}(k,n)$ denote the corresponding eigenvectors. Equipped with these observations, the noisy signal correlation matrix can also be diagonalized as

$$\mathbf{B}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n)\mathbf{B}(k,n) = \mathbf{\Lambda}(k,n) + \mathbf{I}_{MN}. \tag{7}$$

The joint diagonalization can be interpreted as a particular spatiotemporal filterbank decomposition with MN subbands, where the noise is whitened and equalized in all subbands. That is, enhancement filters derived based on such decomposition are robust against non-white noise.

Before moving on to the filter designs, we also introduce the subband input SNR for the considered noise reduction problem:

$$\text{iSNR}(k,n) = \frac{\phi_{X_1}(k,n)}{\phi_{V_1}(k,n)}, \tag{8}$$

where $\phi_{X_1}(k,n) = E\left[|X_1(k,n)|^2\right]$ and $\phi_{V_1}(k,n) = E\left[|V_1(k,n)|^2\right]$ are the variances of $X_1(k,n)$ and $V_1(k,n)$, respectively.

3. VARIABLE SPAN FILTERING

In this section, we then present the variable span filters for noise reduction in the framework presented in Section 2. First, we introduce the STFT domain filtering operation, i.e.,

$$Z(k,n) = \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{y}}(k,n), \tag{9}$$

with $Z(k,n)$ denoting the resulting estimate of the desired signal $X_1(k,n)$. Furthermore, $\underline{\mathbf{h}}(k,n)$ is a complex-valued filter of length MN defined as $\underline{\mathbf{h}}(k,n) = \left[\mathbf{h}^T(k,n) \ \cdots \ \mathbf{h}^T(k,n-N+1)\right]^T$, where $\mathbf{h}(k,n-i)$ is a filter of length M containing all the complex gains applied to the sensor outputs at frequency bin k and time frame $n-i$. We can always write the filter using $\underline{\mathbf{b}}_i(k,n),\ i = 1,2,\ldots,MN$ as basis vectors, since $\mathbf{B}(k,n)$ is full rank, i.e.,

$$\underline{\mathbf{h}}(k,n) = \mathbf{B}(k,n)\underline{\mathbf{a}}(k,n), \tag{10}$$

where $\underline{\mathbf{a}}(k,n) = \begin{bmatrix} A_1(k,n) & A_2(k,n) & \cdots & A_{MN}(k,n)\end{bmatrix}^T$, is the filter representation in the new basis. This means that, instead of finding $\underline{\mathbf{h}}(k,n)$ directly like in conventional approaches, we can equivalently tackle the filter design problem by finding the coordinates $A_i(k,n),\ i = 1,2,\ldots,MN$. If we substitute (10) into (9), we obtain

$$Z(k,n) = \underline{\mathbf{a}}^H(k,n)\mathbf{B}^H(k,n)\left[\underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n)\right]. \tag{11}$$

Using previous assumptions, the variance of $Z(k,n)$ becomes

$$\phi_Z(k,n) = \underline{\mathbf{a}}^H(k,n)\mathbf{\Lambda}(k,n)\underline{\mathbf{a}}(k,n) + \underline{\mathbf{a}}^H(k,n)\underline{\mathbf{a}}(k,n). \tag{12}$$

We then decompose $\underline{\mathbf{a}}(k,n)$ into two subvectors as $\underline{\mathbf{a}}(k,n) = \begin{bmatrix}\mathbf{a}'^T(k,n) & \mathbf{a}''^T(k,n)\end{bmatrix}^T$, where $\mathbf{a}'(k,n)$ is a vector containing the first P coefficients of $\underline{\mathbf{a}}(k,n)$ and $\mathbf{a}''(k,n)$ is a vector containing the last $MN-P$ coefficients of $\underline{\mathbf{a}}(k,n)$. Similarly, we have $\mathbf{B}(k,n) = [\mathbf{B}'(k,n)\ \mathbf{B}''(k,n)]$, and $\mathbf{\Lambda}'(k,n) = \text{diag}[\lambda_1(k,n), \lambda_2(k,n), \ldots, \lambda_P(k,n)]$, where $\mathbf{B}'(k,n)$ is a matrix of size $MN \times P$ containing the first P columns of $\mathbf{B}(k,n)$ and

$\mathbf{B}''(k,n)$ is a matrix of size $MN \times (MN-P)$ containing the last $MN-P$ columns of $\mathbf{B}(k,n)$. From (12) we can see that $\underline{\mathbf{a}}^H(k,n)\underline{\mathbf{a}}(k,n) = \mathbf{a}'^H(k,n)\mathbf{a}'(k,n) + \mathbf{a}''^H(k,n)\mathbf{a}''(k,n)$ represent the residual noise. Intuitively, many optimal noise reduction filters with no more than P constraints should satisfy $\mathbf{a}''(k,n) = \mathbf{0}_{(MN-P)\times 1}$. This enable us to simplify the problem as

$$\begin{aligned} Z(k,n) &= \mathbf{a}'^H(k,n)\mathbf{B}'^H(k,n)\left[\underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n)\right] \\ &= X_{\mathrm{fd}}(k,n) + V_{\mathrm{rn}}(k,n), \end{aligned} \tag{13}$$

where $X_{\mathrm{fd}}(k,n)$ and $V_{\mathrm{rn}}(k,n)$ denote the filtered desired signal and the residual noise, respectively. In other words, we only need to determine $\mathbf{a}'(k,n)$. The variance of $Z(k,n)$ thus becomes

$$\phi_Z(k,n) = \mathbf{a}'^H(k,n)\left[\mathbf{\Lambda}'(k,n) + \mathbf{I}_P\right]\mathbf{a}'(k,n). \tag{14}$$

We can then deduce that the subband output SNR is

$$\mathrm{oSNR}\left[\mathbf{a}'(k,n)\right] = \frac{\mathbf{a}'^H(k,n)\mathbf{\Lambda}'(k,n)\mathbf{a}'(k,n)}{\mathbf{a}'^H(k,n)\mathbf{a}'(k,n)}, \tag{15}$$

and, further derivations reveal that $\mathrm{oSNR}\left[\mathbf{a}'(k,n)\right] \leq \lambda_1(k,n)$.

It is also useful to quantify the distortion introduced by the filter. This can, for example, be measured using the subband desired signal reduction factor defined as

$$\xi_{\mathrm{sr}}\left[\mathbf{a}'(k,n)\right] = \frac{\phi_{X_1}(k,n)}{\sum_{p=1}^{P}\lambda_p(k,n)\left|A_p(k,n)\right|^2}. \tag{16}$$

Obviously, we have no distortion only when $\xi_{\mathrm{sr}}\left[\mathbf{a}'(k,n)\right] = 1$.

If we take a close look on (15), we can see that the subband output SNR is maximized if and only if $A_1(k,n) \neq 0$ and $A_2(k,n) = \cdots = A_P(k,n) = 0$. Consequently, the maximum SNR filter is $\underline{\mathbf{h}}_{\max}(k,n) = A_1(k,n)\underline{\mathbf{b}}_1(k,n)$, where $A_1(k,n) \neq 0$ is an unknown and arbitrary complex number. It can be shown that, if we choose the $A_1(k,n)$ that minimizes the MSE corresponding to distortion,

$$J_{\mathrm{ds}}\left[\mathbf{a}'(k,n)\right] = E\left[\left|X_1(k,n) - \mathbf{a}'^H(k,n)\mathbf{B}'^H(k,n)\underline{\mathbf{x}}(k,n)\right|^2\right],$$

under the assumption of $P=1$, we get

$$\underline{\mathbf{h}}_{\max}(k,n) = \frac{\underline{\mathbf{b}}_1(k,n)\underline{\mathbf{b}}_1^H(k,n)}{\lambda_1(k,n)}\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}, \tag{17}$$

where $\underline{\mathbf{i}}$ is the first column of $\mathbf{I}_{MN}$.

The minimum variance distortionless response (MVDR) filter can also be found in this framework. Minimizing the MSE with respect to distortion, J_{ds}, for an arbitrary P yields the MVDR filter

$$\underline{\mathbf{h}}_{\mathrm{MVDR}}(k,n) = \sum_{p=1}^{P}\frac{\underline{\mathbf{b}}_p(k,n)\underline{\mathbf{b}}_p^H(k,n)}{\lambda_p(k,n)}\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}. \tag{18}$$

From (17) and (18), we see that there is a clear link between this and the maximum SNR filter. Hence, we propose a class of minimum distortion (MD) filters given by

$$\underline{\mathbf{h}}_{\mathrm{MD},Q}(k,n) = \sum_{q=1}^{Q}\frac{\underline{\mathbf{b}}_q(k,n)\underline{\mathbf{b}}_q^H(k,n)}{\lambda_q(k,n)}\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}, \tag{19}$$

where $1 \leq Q \leq P$. The variable Q will control the tradeoff between noise reduction and signal distortion, i.e., higher Q means less noise reduction but also less distortion and vice versa.

We can also derive a Wiener-type filter in this framework, by first introducing an error signal $\mathcal{E}(k,n) = Z(k,n) - X_1(k,n)$ and then the general MSE

$$J\left[\mathbf{a}'(k,n)\right] = E\left[\left|\mathcal{E}(k,n)\right|^2\right] = J_{\mathrm{ds}}\left[\mathbf{a}'(k,n)\right] + J_{\mathrm{rs}}\left[\mathbf{a}'(k,n)\right],$$

where $J_{\mathrm{rs}}\left[\mathbf{a}'(k,n)\right] = \mathbf{a}'^H(k,n)\mathbf{a}'(k,n)$ is the MSE of the residual noise. Minimizing the above MSE yields the Wiener filter:

$$\underline{\mathbf{h}}_{\mathrm{W}}(k,n) = \sum_{p=1}^{P}\frac{\underline{\mathbf{b}}_p(k,n)\underline{\mathbf{b}}_p^H(k,n)}{1+\lambda_p(k,n)}\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}. \tag{20}$$

We can see that the MVDR and Wiener filters are very close to each other; they only differ by the weighting function, which strongly depends on the spatiotemporal subband SNR. The Wiener filter will have an output SNR at least as high as the MVDR filter, but its signal reduction factor will be equal to or higher than that of the MVDR filter.

Finally, tradeoff filters can be obtained by solving

$$\min_{\mathbf{a}'(k,n)} J_{\mathrm{ds}}\left[\mathbf{a}'(k,n)\right] \quad \text{s. t.} \quad J_{\mathrm{rs}}\left[\mathbf{a}'(k,n)\right] = \beta\phi_{V_1}(k,n), \tag{21}$$

where $0 \leq \beta \leq 1$, to ensure that filtering achieves some degree of noise reduction. Solving this yields

$$\underline{\mathbf{h}}_{\mathrm{T},\mu}(k,n) = \sum_{p=1}^{P}\frac{\underline{\mathbf{b}}_p(k,n)\underline{\mathbf{b}}_p^H(k,n)}{\mu+\lambda_p(k,n)}\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}, \tag{22}$$

where $\mu \geq 0$ is a Lagrange multiplier. Inspired by the MD filter design, we introduce the most general tradeoff filter

$$\underline{\mathbf{h}}_{\mathrm{GT},\mu,Q}(k,n) = \sum_{q=1}^{Q}\frac{\underline{\mathbf{b}}_q(k,n)\underline{\mathbf{b}}_q^H(k,n)}{\mu+\lambda_q(k,n)}\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}, \tag{23}$$

where Q does not have to be equal to P. Clearly, all previous filter designs can be obtained using the above tradeoff filter.

4. EXPERIMENTS

We then proceed with the experimental evaluation of the proposed filter designs. For comparison, the multichannel, STFT-domain Wiener filter in [23] was included in the evaluation. We considered a scenario with reverberant speech signals contaminated by diffuse babble noise and white Gaussian sensor noise. The speech signals were two female and two male speech signals from the Keele database [24], amounting to a total of 10 seconds. The speech signals were single-channel, and were therefore synthesized spatially using a room impulse response (RIR) generator [25]. The simulated room had dimensions $3 \times 4 \times 3$ m, and the source was located at $(0.75, 1, 1.5)$ m, while the microphones were placed at $(1.5 + d[m - \frac{M-1}{2}], 2, 1)$ m for $m = 0, \ldots, M-1$ with d denoting the microphone spacing. The sensor spacing was 5 cm and the number of microphones was $M = 3$. Additionally, the speed of sound was 343 m/s, the 60 dB reverberation time was 0.2 s, the room impulse response length was 2,048, and the microphone type was omnidirectional. We then generated our clean, multichannel speech signals including reverberation using this setup. The sensor noise was white Gaussian in each channel, while the diffuse noise was babble noise. To obtain the diffuse babble noise, we used a single-channel babble noise signal from the AURORA database [26] and assumed a spherical noise field. Under this assumption, a multichannel diffuse noise signal can be generated as described in [27]. In all

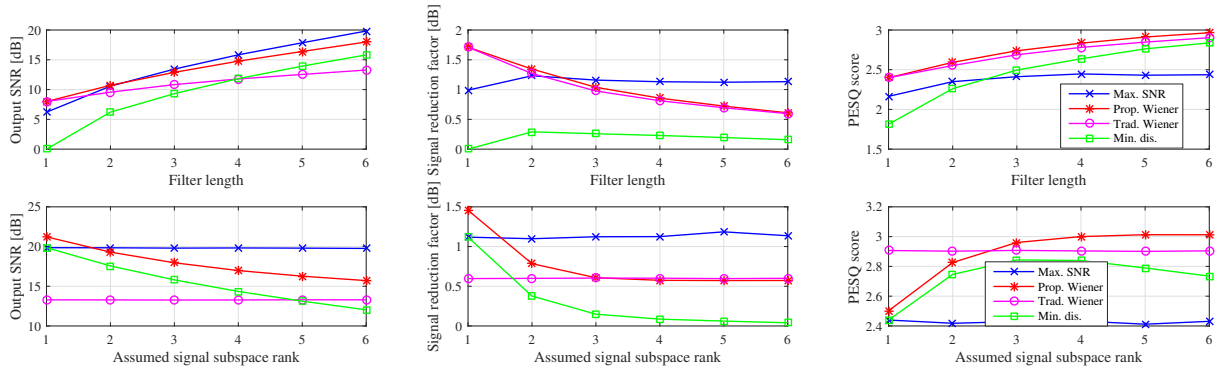

Fig. 1. Filter performances versus (top) the filter length, and (bottom) the assumed signal subspace rank.

evaluations, the sensor and diffuse noise were added at signal-to-sensor-noise ratios and signal-to-diffuse-noise ratios of 30 dB and 0 dB, respectively. Using mixtures of the speech and these different noise types, we then conducted evaluations of the aforementioned filters in terms of their output SNRs, signal reduction factors, and PESQ scores. The PESQ scores were obtained using an implementation (http://ecs.utdallas.edu/loizou/speech/software.htm) of the PESQ standard for speech quality assessment [28], i.e., they are objective but should reflect the subjective quality. In each of the evaluations, enhancement of the speech signal simultaneously mixed with the two noise types were considered, and the performance of the filters were measured over time and averaged.

To compute the coefficients of the proposed filters, we need estimates of the correlation matrices $\mathbf{\Phi}_{\underline{\mathbf{x}}}$ and $\mathbf{\Phi}_{\underline{\mathbf{v}}}$. Estimation of these has not been investigated thoroughly before, and will constitute a research contribution in itself. Since the focus herein is rather on the design of optimal enhancement filters, we therefore estimate the needed statistics directly from the separated signals. We expect, however, that techniques such as VAD [29], minimum statistics [30], sequential methods [31], and multichannel PSD estimators [11, 32] can be generalized for practical statistics estimation. It is important to note that the traditional Wiener also required information about the statistics of the desired signal (or the noise). These have also been estimated from the separated signals in our evaluation to make the comparison fair. The estimation of the correlation matrix of a vector $\underline{\mathbf{a}}(k,n)$ was done recursively as $\widehat{\mathbf{\Phi}}_{\underline{\mathbf{a}}}(k,n) = (1-\xi)\widehat{\mathbf{\Phi}}_{\underline{\mathbf{a}}}(k,n-1) + \xi\underline{\mathbf{a}}(k,n)\underline{\mathbf{a}}^H(k,n)$, where ξ is the forgetting factor, and $\widehat{\mathbf{\Phi}}_{\underline{\mathbf{a}}}(k,n)$ denotes an estimate of $\mathbf{\Phi}_{\underline{\mathbf{a}}}$ at frequency bin, k, and time instance n. The forgetting factors for all statistics estimators in the considered evaluations were 0.05. Furthermore, the STFT's of the signals from the different channels were calculated using rectangularly windowed blocks of 40 samples and an FFT length of 64. The blocks were overlapping by 50 %, and after enhancement, the blocks were combined using overlap-add with Hanning windows.

First, the filter performances were investigated versus the temporal filter length, N, to investigate the benefit of exploiting interframe correlation. We considered a scenario with an assumed signal subspace rank of $Q = 3$, and the simulation setup described above. The results from this evaluation are presented in Figure 1. As expected, the maximum SNR filter has the highest output SNR in most cases, i.e., for filter lengths larger than 2. Since the output SNRs and signal reduction factors are measured from the filter outputs and not using the theoretical expressions in (15) and (16), we can not expect the theoretical relationships to always hold such that the maximum SNR filter always has the highest output SNR. The minimum distortion filter has a somewhat lower output SNR but on the other hand has a lower distortion according to the measured signal reduction factors. Compared to the traditional Wiener filter exploiting interframe correlation (Trad. Wiener), the proposed Wiener filter has a higher output SNR and almost the same amount of distortion. The measured PESQ scores are quite similar to the output SNR measurements, except that the maximum SNR is worse than all other filters for high filter lengths. Most importantly, the proposed Wiener filter outperforms the traditional Wiener also in terms of PESQ score. For higher filter lengths, the difference in PESQ score is around 0.1, which will be audible. Finally, we evaluated the filters for different assumed signal subspace ranks. The filter length was fixed to 6, and, using this setup, we obtained the results in Figure 1. The performance of the maximum SNR and traditional Wiener filters are nearly same for all ranks as expected. Moreover, we see that the output SNRs of the proposed minimum distortion and Wiener filters decrease for an increasing rank, but the signal reduction factor is also lowered at the same time. For all ranks, the proposed Wiener filter has higher output SNR than the traditional one, and their distortion levels are comparable for ranks larger than 2. In terms of PESQ scores, we see that the maximum SNR has the lowest PESQ score for all ranks. Both Wiener filters outperform the minimum distortion filter for all ranks, and for ranks larger than 2, the proposed Wiener filter has a significantly higher PESQ score than the traditional Wiener filter. Our informal listening test confirmed these findings.

5. CONCLUSION

We considered the topic of multichannel speech enhancement and proposed a new class of so-called variable span filters in the STFT domain. These are designed by, first, conducting a joint diagonalization of the correlation matrices of the signal of interest and the noise. The filters are then formed by using the so-obtained eigenvectors as a basis and the eigenvalues as weights. By varying the number of eigenvectors and -values that are included in the filter designs, we obtain a very flexible design with a high degree of control over the amount of noise reduction and signal distortion. In this filter design framework, we proposed maximum SNR, minimum distortion, Wiener, and tradeoff filters. Compared to state-of-the-art subspace methods for speech enhancement, the proposed methods can easily be evaluated in terms of, and optimized for, their output SNRs and signal reduction factors. Our evaluations on real speech data that were spatially synthesized, show that the proposed variable span filters can outperform their traditional counterparts. For example, the Wiener filter in the proposed variable span framework, can outperform the traditional Wiener filter in terms of both output SNR (more than 8 dB improvement) and mean opinion scores (improvement greater than 0.1).

6. REFERENCES

[1] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, Number VII. Springer, 1 edition, 2011.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[6] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[7] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[8] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[9] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[10] J. Benesty, M. Souden, and J. Chen, "A perspective on multichannel noise reduction in the time domain," vol. 74, no. 3, pp. 343–355, Mar. 2013.

[11] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[12] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*, Springer-Verlag, 2009.

[13] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[14] K. Hermus, P. Wambacq, and H. V. hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. on Applied Signal Processing*, vol. 2007, no. 1, pp. 1–15, Sep. 2007.

[15] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, pp. 24, Jun. 2007.

[16] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Process. Lett.*, vol. 9, pp. 204–206, Jul. 2002.

[17] S. M. Nørholm, J. Benesty, J. R. Jensen, and M. G. Christensen, "Single-channel noise reduction using unified joint diagonalization and optimal filtering," *EURASIP J. on Applied Signal Processing*, vol. 2014, no. 1, pp. 37, Mar. 2014.

[18] S. M. Nørholm, J. Benesty, J. R. Jensen, and M. G. Christensen, "Noise reduction in the time domain using joint diagonalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 7058–7062.

[19] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*, vol. 1, Berlin, Germany: Springer-Verlag, 2008.

[20] M. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Berlin, Germany: Springer-Verlag, 2001.

[21] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[22] J. N. Franklin, *Matrix Theory*, Prentice-Hall, 1968.

[23] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, SpringerBriefs in Electrical and Computer Engineering. Springer Science & Business Media, 2011.

[24] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[25] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010, Ver. 2.0.20100920.

[26] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Sep. 2000.

[27] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[28] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," , no. P.862, pp. 1–30, Feb. 2001.

[29] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[30] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[31] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., chapter 4, pp. 91–115. Boston, MA, USA: Kluwer, 2004.

[32] Q. Gong, B. Champagne, and P. Kabal, "Noise power spectral density matrix estimation based on modified IMCRA," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2014, pp. 1389–1395.